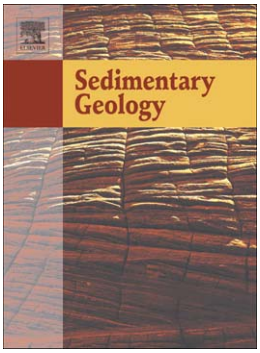

Analyzing coastal environments by means of functional data analysis

Carlos Sierra, Germán Flor-Blanco, Celestino Ordoñez, Germán Flor, José R. Gallego

PII: S0037-0738(17)30143-4
DOI: doi:10.1016/j.sedgeo.2017.06.008
Reference: SEDGEO 5204

To appear in: *Sedimentary Geology*

Received date: 7 April 2017
Revised date: 13 June 2017
Accepted date: 14 June 2017

Please cite this article as: Sierra, Carlos, Flor-Blanco, Germán, Ordoñez, Celestino, Flor, Germán, Gallego, José R., Analyzing coastal environments by means of functional data analysis, *Sedimentary Geology* (2017), doi:10.1016/j.sedgeo.2017.06.008

Analyzing coastal environments by means of functional data analysis

Carlos Sierra[1]*, Germán Flor-Blanco[2], Celestino Ordoñez[3], Germán Flor[2], José R. Gallego[3]

[1]Escuela Politécnica de Ingeniería de Minas y Energía. University of Cantabria. Boulevard Ronda Rufino Peón nº 254. 39316 Torrelavega, Spain

[2]Department of Geology. University of Oviedo. GeoQuo Research Group (Geomorphology and Quaternary). Campus de Llamaquique. C/ Jesús Arias de Velasco, s/n. 33005 Oviedo, Spain

[3]Department of Mining Exploitation and Prospecting, Campus de Mieres, University of Oviedo, 33600 Mieres, Spain

*Corresponding author. E-mail address: carlos.sierra@unican.es.

Abstract

Here we used Functional Data Analysis (FDA) to examine particle-size distributions (PSDs) in a beach/shallow marine sedimentary environment in Gijón Bay (NW Spain). The work involved both Functional Principal Components Analysis (FPCA) and Functional Cluster Analysis (FCA). The grainsize of the sand samples was characterized by means of laser dispersion spectroscopy. Within this framework, FPCA was used as a dimension reduction technique to explore and uncover patterns in grain-size frequency curves. This procedure proved useful to describe variability in the structure of the data set. Moreover, an alternative approach, FCA, was applied to identify clusters and to interpret their spatial distribution. Results obtained with this latter technique were compared with those obtained by means of two vector approaches that combine PCA with CA (Cluster Analysis). The first method, the point density function (PDF), was employed after adapting a log-normal distribution to each PSD and resuming each of the density functions by its mean, sorting, skewness and kurtosis. The second applied a centered-log-ratio (clr) to the original data. PCA was then applied to the transformed data, and finally CA to the retained principal component scores. The study revealed functional data analysis, specifically FPCA and FCA, as a suitable alternative with considerable advantages over traditional vector analysis techniques in sedimentary geology studies.

Keywords: particle-size distribution; sand sediments; functional cluster analysis; functional components analysis; vector-based clusters.

1. Introduction

Particle size and sediment type are the parameters most widely used to characterize sediments (e.g., Krumbein, 1934; McManus, 1988; Poppe et al., 2000). Particlesize determinations are common practice in powder technology, which includes several chemical and industrial activities (Burgess et al., 2004; Allen, 2013). However, such measurements are not limited to industrial purposes alone as they have various applications in the earth sciences (e.g., Sahu, 1964; Weltje and von Eynatten, 2004; Flood et al., 2015).

Several bulk properties of sediments are either dependent on or correlated with ParticleSize Distributions (PSDs), as is case of surface sediments and morphological characteristics (Horowitz and Elrick, 1987), chemical composition (Bloemsma et al., 2012; Tanse and Rafiuddi, 2016), hydraulic and water transport properties (Bear and Verruijt, 2012), and electrical properties (Revil and Jardani, 2013). Geophysical properties, such as compressional and shear wave speeds (Jackson and Richardson, 2007), and geomorphic processes (Lankaster, 1982; Poppe et al., 2000; Vandenberghe, 2013) can also be included in the list.

The best approach to represent the size characteristics of a granular material is through its PSD. This distribution is summarized by tables or plots in the form of frequency (where 'frequency' corresponds to weight or volume of sediment in a particular size class) or of cumulative weight or volume size classes (Syvitski, 2007; Allen, 2013). Using these functions, granulometric parameters can be determined by graphical methods, which are the most widely used approaches for this purpose (Folk and Ward, 1957), or by mathematical methods (Blott and Pye, 2001).

Such granulometric parameters include those obtained from cumulative frequency functions, namely the first (mean, Mz), second (standard deviation, SD), third (skewness, SK), or fourth (kurtosis, K) moments, or from cumulative functions, such as quartiles (Q_1, Q_2, Q_3, Q_4) (e.g., the QDa-Md method) (Buller and McManus, 1972; Duck, 1994) and decile diameters (e.g., D_{10}, D_{25}, D_{75}) (Merkus, 2009; Allen, 2013). Moments are usually calculated by fitting normal, lognormal (Blott and Pye, 2001) or even a log-hyperbolic and log-skewed distribution (e.g., Barndorff-Nielsen and Christiansen, 1988; Weltje and Prins, 2007; Donato et al., 2009).

The aforementioned statistical moments can be useful for the following: i) characterizing depositional sub-environments (e.g., Passega, 1957; Friedman, 1967; Solohub and Klovan, 1970; Martins, 2003; Reineck and Singh, 2012); ii) evaluating aeolian and beach environment interactions (Knight et al., 2002; Weltje and von Eynatten, 2007). However, differentiation between beach and dune sands is not straightforward because the transport distance from beach to foredunes is short and beach sediments are usually well sorted (e.g., Arens et al., 2002); iii) differentiating selective erosive and depositional periods in foreshore environments (e.g., McLaren et al., 2007); iv) interpreting transport hydrodynamics (e.g., Kairyté and Stevens, 2014; Flood et al., 2015) in bays, embayed beaches as a regular grid (Gao and Collins, 1992, 1994), or along sand drifts in low-coastal beaches (McCave, 1978); and v) establishing the differences between the beach, an associated aeolian dune field and the sand transfers between them (Arens, 1994).

In this context, functional data analysis (FDA) can provide a more suitable approach to work with grain-size data than classical multivariate analysis (MA), the latter originally developed to deal with vector data and independent covariates (Ramsay and Silverman,

2000). FDA addresses the same type of questions as MA, such as regression, supervised classification, and clustering or principal components among others (Celeux and Govaert, 1995). A variant of MA is the well-known end-member modeling analysis (EMMA) based on the hypothesis that each PSD is a non-negative linear combination of fundamental components, called end-members, and extracted from the PSD (Weltje, 1997; Dietze et al., 2012). However, in FDA it is assumed that the underlying curve-generation process is smooth and that the data measured are dependent. Therefore in FDA the raw data are extracted for a continuous smooth function, so it is possible to estimate not only undamped data but also the derivatives of the functions (Fraley and Raftery, 2002), which can be useful for some types of analysis. Finally, FDA works with functions instead of scalars.

There are two main techniques among FDA. The first, Functional Cluster Analysis (FCA), is a mixture model-based clustering method where PSDs are assumed to have been generated by a finite mixture of K Gaussian distributions, K being the number of clusters (Gaffney and Smith, 1999). The second, Functional Principal Components Analysis (FPCA), is an approach where each component contains information about the characteristics of the curves that are associated with the clusters (Ramsay and Silverman, 2000; Febrero-Bande and Oviedo, 2012).

All things considered, the aim of the present study is to introduce FDA to the sedimentology community. This is done by means of two novel approaches in sediment studies, namely FPCA and FCA. The results obtained were compared with traditional vector-based procedures. Both mathematical procedures were applied to a data set of PSDs with small variability, that is to say, of similar size classes (coarse to very fine sand) and comparable PSDs shapes in which apportionment was not easy.

2. Geographical setting

The study area is located in the city of Gijón (Asturias, northern Spain), in the central Asturian mesotidal cliff coast of the Cantabrian Sea (Fig. 1). Gijón has an important industrial area in its surroundings and therefore the environmental compartments (including sediments) require frequent monitoring (Boente et al., 2017). In this context, the city has three urban embayed sandy beaches (Fig. 1), San Lorenzo, located to the east, being the broadest. The peninsula of Santa Catalina (which has a small yacht port on the left side) separates this beach from two artificial beaches, namely, Poniente and El Arbeyal (Flor et al., 2008), which were built in 1994. The Piles coastal river flows directly into the eastern side of San Lorenzo beach, forming the main channel and inlet of an old small estuary. Sedimentary studies in the area have been done in the last years as a result of the erosion caused in this important touristic and industrial area by the expansion works in the port of El Musel. Moreover, a recentincrease of erosion caused by storms as well as by sea level rise has also been reported (e.g, Flor-Blanco et al., 2016). In this respect, beach nourishment with sand dragged seawards the closure depth, was proposed as possible solution.

This coast has a temperate oceanic and humid climate (average monthly minimum of 5ºC and maximum of 24ºC). It registered an average monthly minimum rainfall of 60.7 mm and maximum of 125.8 mm (average of 956.3 mm/year) in the period 2000 to 2012 (World Weather Online, 2017). Astronomical tides (about 68.3%) are mesotidal (Flor-Blanco et al., 2015) and semidiurnal (Gómez et al., 2014). Typical spring-neap cycle tidal ranges vary between approximately 4.8 m (extreme spring) and 0.7 m (extreme neap). The average tidal range is 2.47 m and the maximum 4.96 m (Ministerio de Fomento, 2010).

On the coast of Asturias, the winds blow from the SW, but NW, W and NE winds directly affect the dynamics of the beaches and are responsible for the formation of dunes (Flor and Flor-Blanco, 2014). The winds in Gijón are SW (11% of the time) and W (12%), while E (11%) and NE (6%) winds occur under cyclonic atmospheric conditions or high-pressure, followed by NW (8%) winds and those from other directions. In the case of the study area, Cape Torres protects it from NW and W winds. Calm conditions account for about 16% of all wind events (García et al., 2010).

The main frequencies of wave direction are from the fourth quadrant (WNW-NW) (Supplementary Material, SM1), and to a lesser extent, the first quadrant (N and NE). The average wave height is 1.0 m, and stormy conditions occur when wave heights exceed 2.0 m. Under storm waves, with a periodicity of 100 years, the average wave height is 9.75 m, while maximum wave height and length are 17.55 m and 415 m, respectively.

Closure depth was calculated only for San Lorenzo beach because Poniente beach and Arbeyal beach are artificial and their sand prism is not deeper than 1.5 m. According to the data from the Buoy of Gijón, closure depth ranges between 9.23 m and 11.02 m below sea level (ROM_0.3-91). When biological criterion of the polychaete tube worm (*Galathowenia oculata)* was applied, closure depth ranged between 12 and 15 m. Bathymetric data are presented in the supplementary materials section SM2, wherein the narrow belt between bottoms (convex isobaths seawards vs. concave isobaths landwards) can be appreciated at an estimated depth of 12 m.

Sands in these beaches have a high proportion of carbonate bioclasts and quartz. Bioclasts are supplied by a variety of organisms with calcareous shells inhabiting the stretch between the supratidal and subtidal zone. The bay is affected by a local

upwelling, where the coastal current moves towards the E (Flor, 1981), thereby increasing nutrient content and organism communities in cliffs. Siliciclastic sand fractions come from Nalón River, located 41 km to the west, and probably to a lesser extent from the Piles River basin. Transgressive sedimentary infill is formed by mixed sands in a coastal area where upwelling allowed the formation of bioclastic sands and organic muds (Flor and Lharti, 2008).

The embayed beach of San Lorenzo is linked to an old estuary. This is now reduced to a narrow artificial channel of Piles River that is affected by tides and river discharges. The beach is in dynamic equilibrium during fair weather episodes, mainly as a consequence of waves coming from the NW and NE. It has a highly energetic zone in most of its center which extends eastwards. Sand transport in this area is induced by the longshore eastward currents that arise from the residual component created by the oblique incoming NW waves. In the eastern corner of the beach, the rocky bottom outcrops due to river discharges and ebb tide water, thereby hindering beach-drift sedimentation. The western side of San Lorenzo has a zone of wave shadow (sheltered area caused by the refraction and diffraction of the waves on Santa Catalina Headland). A preliminary model of the beach dynamics is presented in supplementary material SM3.

3. Materials and methods

3.1 Sampling survey

A total of 75 samples was collected, of which 28 corresponded to the natural embayed beach of San Lorenzo sampled until the calculated depth of closure (Flor-Blanco et al., 2016), 30 to the offshore belt, 10 to surface sands from Poniente beach, and 7 to surface sands from El Arbeyal beach (Fig. 1). Beach samples aim to be representative of each

morphological unit (e.g. berm, backshore, foreshore, ripples area, ridge and runnel). In submerged areas, samples were collected using an approximate sampling grid.

The emerged samples were collected randomly during spring low tides, while submerged ones were taken with a Petersen dredge sampler from a boat equipped with sonar, GPS and Hypack hydrographic survey software. These samples corresponded to sheltered area from wave energy, equilibrium morphosedimentary state and equilibrium beach profile (Dean, 1991). In order to consider variability in both long and short distances, submerged samples were collected from consecutive points at an average distance of 200 m and a minimum distance of around 10 m. Due to the high variability of the sand from the distinct beaches, the sampling campaign reduced the approximate distance between samples to 5 m, each sample representing a geomorphological zone of the beach for each profile.

3.2 Sediment characterization

Samples, 200 g in weight, were dried at 30ºC, and then organic matter was removed with hydrogen peroxide (Gee and Bauder, 1996). Subsequently, approximately 20 g representative subsamples, which were used for particlesize characterization, were disaggregated and dispersed with sodium hexametaphosphate and sodium carbonate.

Carbonate concentrations were determined by HCl acid attack and difference in weight measurement. The percentage of carbonate indicates the percentage of biogenic versus siliciclastic grains. This parameter is generally related to the mean, thus the greater the carbonate content, the coarser the grain size.

Particle-size characterization (grain-size range 0.02 - 2000 μm) was performed by means of laser diffraction spectroscopy in the Aqueous Liquid Module of a LS 13-320

MW model (Beckman Inc. Coulter), following the procedure described in ISO 13320 (2009) and adopting the Fraunhofer theories of light scattering. Descriptive parameters were calculated by means of GRADISTAT (Blott and Pye, 2001) according to Folk and Ward (1957), which includes the most sensitive parameters, namely the mean, hydraulic sorting, skewness, and kurtosis, to distinguish between sub environments.

The coarsest one percentile value represents the greatest kinetic energy of deposits. Mean (Mz) and median grain size (D_{50}) are considered indicators of the average kinetic energy during transport and sedimentation (Passega, 1957; Blair and McPherson, 1999). Hydraulic sorting (σ_I) represents the dispersion of the sediment size distribution and indicates the selective sedimentation of the material, which is in movement. Sorting is poor when the proportion of coarser particles decreases (Folk and Ward, 1957). Skewness measures the degree of asymmetry of a distribution function. Thus, positive skewness indicates a high content of fine material and deposition from material in suspension. In contrast, negative skewness suggests an excess of coarse material and also transport on the bottom through saltation (Opreanu et al., 2007). Kurtosis is a measure of the "tailedness" of a statistical distribution. When the central portion is better sorted than the tails, the curve is leptokurtic; whereas when the tails are better sorted than the central portion then the distribution is flat-peaked and therefore platykurtic (Friedman and Sanders, 1978).

3.3 Functional principal components analysis (FPCA)

Like its counterpart in standard multivariate analysis (Principal Components Analysis, PCA), FPCA is often used to interpret the underlying structures of a dataset (Ramsay and Silverman, 2000). In this regard, FPCA is used to obtain the patterns encoding the highest variance in a data set and not to perform direct maximization of the separation

between groups of samples (Ramsay and Silverman, 2000). FPCA groups variables instead of observations; however, it can act in a similar manner to PCA since directions with the most variance in some data sets are those represented by the first principal components (Bali et al., 2014). In these cases, the first components are the most relevant for clustering because they comprise the most dominant patterns. Other uses of FPCA include outlier detection and quality control (Sawant et al., 2012).

As in PCA, in FPCA we look for the eigenvalues and eigenvectors of the covariance matrix. This leads to the following eigenequation:

$$\int \Sigma(s,t)\xi(t)dt = \lambda\xi(s) \tag{1}$$

where the covariance function Σ is given by

$$\Sigma(s,t) = \frac{1}{n}\sum_{i=1}^{n} y_i(s)\, y_i(t) \tag{2}$$

and ξ represents the weight functions (scores of the PCA) and λ an eigenvalue.

Equation (1) can be reduced to a matrix form by means of an expansion on $m<n_i$ basis functions ϕ of each function y_i, as follows:

$$y_i(t) = \sum_{j=1}^{m} c_{ij}\phi_j(t) \tag{3}$$

Similarly, the eigenfunction ξ has an expansion

$$\xi(s) = \sum_{j=1}^{m} b_j\phi_j(t) \tag{4}$$

The basis function expansion of the observed functions, $\mathbf{y} = (y_1,...,y_n)'$, and weights, $\xi = (\xi_1,...,\xi_n);\ \|\xi\| = 1$, can be written more compactly, in matrix form, as

$$y = \mathbf{C}\phi; \quad \xi = \phi'\mathbf{b} \tag{5}$$

Where C is a $n \times m$ matrix of coefficients, **b** a coefficient vector of dimension $m \times 1$, and $\phi' = (\phi_1,...,\phi_m)$ a column vector of basis functions.

Then, the term on the left in equation (1) can be written as follows:

$$\begin{aligned} &\int \Sigma(s,t)\xi(t) = \frac{1}{n}\int \phi(s)'\mathbf{C}'\mathbf{C}\phi(t)\phi(t)'\mathbf{b}dt = \frac{1}{n}\phi(s)'\mathbf{C}'\mathbf{C}\left(\int \phi(t)\phi(t)'dt\right)\mathbf{b} \\ &= \frac{1}{n}\phi(s)'\mathbf{C}'\mathbf{C}\mathbf{W}\mathbf{b} \end{aligned} \tag{6}$$

where $\mathbf{W} = \int \phi(t)\phi(t)'\,dt$.

W=**I** for orthonormal basis functions, like Fourier series. In other cases, it has to be calculated by numerical integration.

By substituting (6) in (1), we obtain the eigenequation in matrix form:

$$\frac{1}{n}\mathbf{C}'\mathbf{C}\mathbf{W}\mathbf{b} = \lambda\mathbf{b} \tag{7}$$

By solving (7) by the least-squares method, the coefficient vector **b** and the eigenvalues λ are estimated. The eigenfunctions are then obtained from (4). When **W**=**I**, the functional PCA problems reduce to a standard PCA on the array **C.**

3.4 Curve clustering model

To obtain information related to the origin and spatial distribution of the samples, we built a mixture model-based cluster, following Gaffney (2004), to group the PSD data.

Let $\mathbf{Y}=\{y_1,...,y_n\}$ be a sequence of curves observed at $\mathbf{X}=\{\mathbf{x}_1,...,\mathbf{x}_n\}$, respectively. In general, the number of observations in each curve n_i can differ. In this model, the curves are considered to be realizations of *n* independent and identically distributed random variables and are assumed to have been generated by a finite mixture of *K* probability density functions. The conditional probabilistic cluster model that relates y_i to $\mathbf{x}_i$ can be expressed as follows:

$$p\left(y_i|\mathbf{x}_i,\Theta\right)=\sum_{k=1}^{K}\alpha_k p_k\left(y_i|\mathbf{x}_i,\boldsymbol{\theta}_k\right) \tag{8}$$

where $p_k(\cdot)$ represents each of the density functions associated with each cluster, $\boldsymbol{\theta}_k$ a vector with the parameters of the density function , and α_k non-negative mixture weights that sum one. Each curve y_i, *i* =1,…,*n*, is considered to be generated by only one cluster.

In our particular model, the conditional probability density functions are obtained from a polynomial regression model, as follows:

$$y_i=\mathbf{X}_i\boldsymbol{\beta}+\varepsilon_i,\qquad \boldsymbol{\varepsilon}_i\quad \mathcal{N}\left(\mathbf{0},\sigma^2\mathbf{I}\right) \tag{9}$$

where $\mathbf{X}_i$ represents a $n_i\times(p+1)$ matrix with the particle diameters, $\boldsymbol{\beta}$ a (p+1)-vector of regression coefficients and ε_i an error term. By considering the data as curves instead of vectors, we can take advantage of the smoothness of the continuous information contained in the sequence of data that is not explicit in vector form.

By combining equations (8) and (9), we can rewrite the cluster model as a mixture of Gaussian functions:

$$p\left(y_i \middle| \mathbf{x}_i, \Theta\right) = \sum_{k=1}^{K} \alpha_k \mathcal{N}\left(y_i \middle| \mathbf{X}_i \beta_k, \sigma_k^2 \mathbf{I}\right) \tag{10}$$

where $\Theta = \left(\alpha_1, ..., \alpha_k, \beta_1, ..., \beta_k, \sigma_1^2, ..., \sigma_k^2\right)$ is the complete parameter vector.

The model in (10) can be reduced to a problem of finding the maximum of the log-likelihood function:

$$\mathcal{L}_c = \sum_{i=1}^{n} \log \alpha_{z_i} \mathcal{N}\left(y_i \middle| \mathbf{X}_i \beta_{z_i}, \sigma_{z_i}^2 \mathbf{I}\right) \tag{11}$$

z_i being the unknown cluster membership for curve y_i

The estimation of maximum likelihood parameters in (11) can be obtained by means of the EM (Expectation-Maximization) algorithm (Dempster et al., 1997; McLachlan and Krishnan, 1997; Gaffney et al., 2007). This iterative algorithm is called EM because it has two steps: the E-step (expectation) and the M-step (maximization). In the former, the conditional expectation of the log-likelihood is computed on the basis of the observed data and the current parameter estimates, while the latter determines the M-step parameters $\left\{\beta_k, \sigma_k^2 . \alpha_k\right\}$ that maximize the expected log-likelihood from the E-step.

As the number of clusters K contained in the data is not known, this value must be estimated. Several methods can be used for this purpose, such as the Silhouette (Rouseeuw, 1986), the Gap (Tibshirani et al., 2001), the Akaike Information Criterion (AIC) (Akaikem, 1974), and the Bayesian Information Criterion (BIC) (Schwarz, 1978). Here we used the BIC to estimate K because it penalizes the number of parameters more than AIC. The BIC score is given by the following:

$$BIC = \mathcal{L}_{\max} - \frac{1}{2} f \ln n \tag{12}$$

where $\mathcal{L}_{max}$ is the maximum likelihood obtained from the EM algorithm, and the second term is a penalty factor to consider, as the first term increases with the number of independent parameters given by $f = (k-1)+kp+kp(p-1)/2$.

We estimated the BIC for various values of k, the optimum value being k = K, which provides the maximum value of BIC.

3.5 Vector-based methods

The results obtained with FCA were compared with another two methods that consider data as vectors instead of functions. The first method, the point density function (PDF) approach, consists of adjusting a log-normal distribution to each PSD, characterizing each of the density functions by its mean, sorting, skewness and kurtosis (Krumbein and Pettitjohn, 1932). PCA is performed on these parameters, and finally, a vector clustering method (we specifically applied Ward's method) groups the data using the retained principal component scores. In the second approach, we followed the methodology successfully implemented by Flood et al. (2015) to other sediment datasets. In this regard, we first transformed the original compositional data into a new unconstrained set by applying a centred-log-ratio (clr) transformation (Aitchison, 1986). A PCA was then applied to the transformed data, and finally, as in the previous method, Cluster Analysis (CA) (again, by means of Ward's method) was applied to the retained principal component scores.

4. Results and discussion

4.1 Data characteristics

Figure 2 shows the frequency curves for all the PSDs included in the study. The probability density functions evidenced variations in symmetry, ranging from positive

to negative, with several samples being nearly symmetrical. This high variability can be contextualized by the distinct sub-environments, namely the backshore, foreshore, nearshore and offshore areas of San Lorenzo, and the backshore of Poniente and El Arbeyal.

We observed a clear prevalence of unimodal distributions, thereby suggesting relatively low mixing but high sorting. Mean grain sizes ranged between 62.5 and 550 μm. This high variation indicates disparity in the kinetic energy conditions during the deposition of coarse sands (e.g., Blatt et al., 1972; McCave and Syvitski, 1991). Thus, offshore sites generally had smaller mean grain sizes than those of beach areas. We detected a significant difference derived from the closing depth limit, such that size classes varied offshore from fine to very fine sand (9 - 180 μm) and on the beach from fine to medium grained sand (180 - 550 μm).

Moreover, sorting was found to be consistent, with most samples being classified as moderately sorted, with some well sorted ones in the El Musel port zone. Regarding skewness, this parameter ranged from positive to very positive, with no spatial trends. Concerning kurtosis, in the backshore and the shallow intertidal area of San Lorenzo beach, samples were mainly platykurtic and mesokurtic, whereas in the remaining areas they were principally leptokurtic.

Percentile and carbonate content were directly related to the mean grain sizes, and their maximum values were located in high-energy zones, that is to say, the area between the beach and the closure depth, and also in those sectors where grain size fluctuated between 2000 to 700 μm (largest sizes). In the case of carbonates, with few exceptions, the samples containing between 50% and 73% were found in the strip below the closing depth. In the offshore area, the carbonate content was below 40%.

4.2 Functional Principal Components Analysis (FPCA)

Visual inspection of Figure 2 does not provide clear grainsize separations. Within this context, a mathematical tool capable of apportioning the sediments while at the same time uncovering part of the underlying sedimentary conditions would be advisable. In this regard, we recommend that FDA be used on particle-size frequency curves.

Figure 3 depicts the four smoothed principal component functions for the data set. All together, they explained 94.5% of the data variability, of which 74.9% was explained by the first and 9.6% by the second components. It should be taken into account that in order to obtain interpretable kinds of function in terms of the original data, we have to overlook the compositional character of the data. For ease of interpretation, we have therefore sacrificed mathematical rigor and assumed possible errors in the results (see Delicado, 2011, for an analysis of the problems associated with FDA applied to density functions).

Thus, the sign of FPC1 was negative, falling between around 50 μm and 280 μm (from coarse-silt to medium sand, that is, the smallest sediments of the data set), being positive from the coarser grain size onwards. Therefore, FPC1 discriminates between finest sediments and the rest of the particles in the sample. Moreover, FPC2 also showed a change of sign along the grain size. Thus, positive values of FPC2 were between 50 μm and almost 200 μm and they corresponded to the grain-size interval where the PSDs of the medium size particles showed a positive steep slope (note that they have positive sign on FPC1). Moreover, the negative part of the curve, between approximately 200 and 600 μm, corresponded to the coarsest sediments, showing a positive steep slope (these sediments have positive values on FPC1). Therefore, it can be concluded FPC2 discriminates between medium and coarse particles.

Figure 4 (a) and Table 1 indicate the granulometric curves with predominantly negative scores (discontinuous line) and positive scores (continuous line) of the first two FPCs. In addition, average parameters of the positive and negative values of the first two principal components are shown in Table 1. Predominantly negative values of FPC1 corresponded to fine, well sorted, symmetrical and mesokurtic offshore sand sediments. In addition, predominantly positive values were related to coarse, moderately sorted, symmetrical, and mesokurtic sand samples.

Figure 4 (b) makes the analogous consideration for FPC2. Within this context, negative values were associated with medium grained, moderated sorted, negatively skewed and mesokurtic sand samples. Positive values of the second component corresponded to coarse, moderately sorted, symmetrical, positively skewed in most cases and generally mesokurtic sand samples. Two figures summarizing these conclusions are provided in Supplementary Material SM4.

In summary, negative FPC1 scores were more frequent for San Lorenzo beach, while positive FPC2 scores prevailed for El Arbeyal beach and Poniente beach. These results indicate that FPCA appears to separate sediments on the basis of genesis and transport conditions, regardless of their spatial distribution.

4.3 Functional cluster analysis (FCA)

FCA was used to determine sedimentary relations based on PSD associations. Figure 5 shows the score according to the BIC (Bayesian information criterion) for different numbers of clusters. From this graph, it can be observed that the BIC score increases up to the knee point, which is located at 3-4 groups. In this respect, and given the BIC, the FPCA, and the difficulties to interpret four clusters, we examined three clusters;

however, results for four clusters are provided in the supplementary material section SM5.

From Figure 6, it is possible to conclude that all curves partitioned into the same cluster share a common shape and thus similar statistical descriptors. Table 2 provides a summary of statistical descriptors of each of the clusters obtained. Mean values of around 250 μm (ranging from fine to very fine sand) corresponded to cluster 1. Moreover, sizes of roughly 420 μm (medium sands) were predominant in cluster 2. Finally, mean grain size values of approximately 670 μm (coarse sand) were characteristic of cluster 3.

As regards carbonates, concentrations of < 55% corresponded to cluster 1, 55 - 70 % to cluster 2, and 70 - 75% to cluster 3. Carbonate content was found to be well correlated with the centile, so that values between 1000 μm φ and 1250 μm (coarse to very coarse sands), corresponded to cluster 3. Moreover, values between 1000 μm and 700-1000 μm (medium to coarse sands) were grouped mainly in cluster 2. Finally, sizes <700 μm (the smallest) were classified as cluster 1.

Concerning sorting, values of 250 μm (moderately sorted to well sorted) were prevalent in cluster 1, of around 260 μm (moderately well sorted) in cluster 2, and of 370 μm (moderately sorted) in cluster 3. Regarding kurtosis, the predominant value in cluster 1 is 3 (mesokurtic), 7 in cluster 2 (leptokurtic), and 3 in cluster 3 (mesokurtic). Moreover, and unlike the rest of the parameters, skewness did not appear to contribute to sediment apportionment, as determined by FCA.

Regarding spatial distribution, cluster 1 was well represented seawards of the closure depth of San Lorenzo beach and corresponded to the deep areas of the submerged beach and the nearshore modern sand prism. Cluster 3 grouped the backshore samples of El

Arbeyal and Poniente. Almost the entire intertidal beach of San Lorenzo corresponded to cluster 2 and only the upper very narrow supratidal beach on its eastern side (generally as a berm) was classified as cluster 3 (Fig. 7). These results are coherent slightly modifying the initial conceptual model presented in supplementary material SM3.

The PSDs of surficial sediments in each facies varied as a result of the PSDs of parent material, selective and destructive transport processes, and the hydrodynamic characteristics of deposition (Friedman and Sanders, 1978). Accordingly, samples closer to each other should have a higher degree of similarity, although this premise may be contravened by the presence of placers, on which larger particles accumulate; surficial sedimentary structures; shells, wood and litter fragments; and the coastal strip. Despite these considerations, the procedure clearly grouped most of the frequency curves on the basis of a similar depositional environment and transport mode and dynamics (Fig. 7).

When the spatial distribution of the samples is analyzed on the basis of their association in clusters, as well as on the sign of the scores in FPC1 and FPC2, it can be appreciated that those with high positive scores in FPC1 coincided with samples in cluster 3. Moreover, samples with high negative scores in FPC1 were grouped in cluster 1. Likewise, samples with high negative values in FPC2 were grouped mainly in cluster 2, while those with high positive scores in FPC2 were found in cluster 3.

Figure 8 (a) shows the PSD assigned to each cluster following the PDF approach. Comparing it with Figure 6, we can observe that similar results were obtained using FCA and this method. In fact, only four curves were classified in different groups. For instance, looking in detail at Figure 8 (a), it is possible to appreciate a blue curve inside

the group of red curves, possibly indicating that it is erroneously grouped. Something similar occurs with a red curve at the bottom of Figure 8 (a), which should probably be assigned to the green cluster.

Table 3 shows the loadings of the PCA. As can be seen, the mean is the parameter that provides almost all the information in the first principal component. Given that the first principal component explains 99% of the variance, we conclude that almost all the information in this principal component comes from the mean of the PSD.

The second vector-based method provided significantly different results from FCA (also to those of the PDF approach), as it can be observed by comparing Figures 6 and 8 (b). Many of the original grain-size curves were assigned to only one heterogeneous cluster, and the other two clusters were also heterogeneous. Therefore, this method was clearly unsuccessful.

One reason for the differences observed in the results may be consequence of the fact that PSDs do not actually follow a log-normal distribution (e.g., Friedman, 1962; Bagnold and Barndorff-Nielsen, 1980). Moreover, similar distribution coefficients derived from a log-normal function may be obtained from remarkably different PSDs (e.g., Roberson and Weltje, 2014; Flood et al., 2015). Therefore using traditional parameters derived from the log-normal, or even from any other distribution, since not all samples in large datasets follow the same distribution, may result in data misrepresentation. In this respect, it seemed as if FDA could better conceal the true variability present in the data.

5. Conclusions

Here we applied functional data techniques to examine the spatial distribution of sediment particles in a coastal environment represented by three embayed sand beaches.

In this regard, we tested two functional data procedures, namely FPCA and FCA.

With respect to FPCA, a dimension reduction technique, permitted identification of the grain-size fractions of the data set to which greater variations could be attributed. Moreover, this procedure correctly captured the natural characteristics of the data, providing geologically interpretable results.

As regards cluster analysis by FCA, it allowed the identification of compact, well defined and geologically significant clusters. Thus, the procedure correctly summarized and compressed grain-size data, allowing the labeling of grain-size categories on the basis of their relationships. In this respect, three clusters were observed, namely cluster 1, which was the most widely represented in deep areas, cluster 2, which corresponded mostly to the small intertidal beaches of Poniente and El Arbeyal and the large upper tidal beach, and cluster 3, which corresponded to shallow submerged beaches and the entire intertidal beach of San Lorenzo.

From a comparative point of view, FCA provided more interpretable results than those obtained from FPCA. Thus, significant differences with FCA were observed in what refers to cluster sizes, compactness, separation degree, inner similarity, and interpretability of the results. Nevertheless, FPCA continues to be relevant in preliminary apportionments, cluster number determinations, and outlier detection.

In summary, our results support the potential of FDA as an alternative to vector methods such as PCA and CA, which are traditionally used in sedimentology. Thus, we propose that our procedure could be extended to the identification of regular environments and sub-environments, environment interaction studies, as well as transport direction identification.

Acknowledgments

The authors thank the authorities of the Principality of Asturias and the Municipality of Gijon for support received during the course of this research.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall.

Allen, T., 2013. Particle size measurement. Springer, Berlin, Germany.

Álvarez Cabal, A., 2012. Estudio de la variabilidad estacional de la playa arenosa de San Lorenzo (Gijón, Asturias). Master Thesis (unpublished). Departamento de Geología, Universidad de Oviedo, Oviedo. Spain.

Akaike. H., 1974. A new look at statistical model identification. IEEE Transactions on Automatic Control 19, 716-723.

Arens, S.M., 1994. Aeolian processes in the Dutch foredunes. PhD Thesis (unpublished). University of Amsterdam, Amsterdam, The Netherlands.

Arens, S.M., van Boxel, J.H., Abuodhas, J.O.Z., 2002. Changes in grain size of sand in transport over a foredune. Earth Surface Processes and Landforms 27, 1163-1175.

Barndorff-Nielsen, O.E., Christiansen, C., 1988. Erosion, deposition and size distributions of sand. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences 417, 335-352.

Bali, J.L., Boente, G., 2014. Robust functional principal component analysis. In New Advances in Statistical Modeling and Applications. In: Pacheco, A., Santos, R., Oliveira, M.R., Paulino, C.D. (Eds). Springer International Publishing, Berlin, Germany, pp. 41-54.

Bear, J., Verruijt, A., 2012. Modeling groundwater flow and pollution (Vol. 2). Springer Science & Business Media Berlin, Germany.

Blair, T.C., McPherson, J.G., 1999. Grain-size and textural classification of coarse sedimentary particles. Journal of Sedimentary Research 69, 6-19.

Bloemsma, M.R., Zabel, M., Stuut, J.B.W., Tjallingii, R., Collins, J.A., Weltje, G.J., 2012. Modelling the joint variability of grain size and chemical composition in sediments. Sedimentary Geology 280, 135-148.

Blott, S.J., Pye, K., 2001. GRADISTAT: a grain size distribution and statistics package for the analysis of unconsolidated sediments. Earth Surface Processes and Landforms 26, 1237-1248.

Boente, C., Matanzas, N., García-González, N., Rodríguez-Valdés, E., Gallego, J. R. 2017. Trace elements of concern affecting urban agriculture in industrialized areas: A multivariate approach. Chemosphere, 183, 546-556.

Burgess, J., Duffy, E., Etzler, F., Hickey, A., 2004. Particle size analysis. AAPS Workshop Report, Cosponsored by the Food and Drug Administration and the United States Pharmacopeia. AAPS Journal 6, 23-34.

Celeux, G., Govaert, G., 1995. Gaussian parsimonious clustering models. Pattern Recognition 28, 781-793.

Christiansen, C., Hartmann, D., 1988. On using the log-hyperbolic distribution to describe the textural characteristics of eolian sediments—discussion. Journal of Sedimentary Petrology 58, 159-160.

Dean, R.G., 1991. Equilibrium beach profiles: characteristics and application. Journal of Coastal Research 7, 53-84.

Delicado, P., 2011. Dimensionality reduction when data are density functions. Computational Statistics and Data Analysis 55, 401 – 420.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1997. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 39, 1-38.

Dietze, E., Hartmann, K., Diekmann, B., IJmker, J., Lehmkuhl, F., Opitz, S., Stauch, G., Wünnemann, B., Borchers, A., 2012. An end-member algorithm for deciphering modern detrital processes from lake sediments of Lake Donggi Cona, NE Tibetan Plateau, China. Sedimentary Geology 243, 169-180.

Donato, S.V., Reinhardt, E.G., Boyce, J.I., Pilarczyk, J.E., Jupp, B.P., 2009. Particle-size distribution of inferred tsunami deposits in Sur Lagoon, Sultanate of Oman. Marine Geology 257, 54-64.

Febrero-Bande, M., Oviedo de la Fuente, M., 2012. Statistical computing in functional data analysis: The R Package, Journal of Statistical Software 51, 1-28.

Flood, R.P., Orford, J.D., McKinley, J.M., Roberson, S., 2015. Effective grain size distribution analysis for interpretation of tidal–deltaic facies: West Bengal Sundarbans. Sedimentary Geology 318, 58-74.

Flor, G., Lharti, S., 2008. Estratigrafía y sedimentología del recubrimiento costero de la ciudad de Gijón (Asturias). Trabajos de Geología 28, 137-157.

Flor, G., Flor-Blanco, G., 2014. Componentes de viento generadores de morfologías y campos de dunas costeras en Asturias (NO de España). Cuaternario y Geomorfología 28, 47-68.

Flor, G., Flor-Blanco, G. Escribano, R., 2008. Características morfológicas y sedimentarias de la playa artificial de Poniente (Gijón, Asturias). Evolución 1995-2004. Territoris, Universidad Illes Balears 7, 145-156.

Flor-Blanco, G., Flor, G., Pando, L., 2013. Evolution of the Salinas-El Espartal and Xagó beach/dune systems in north-western Spain over recent decades: evidence for

responses to natural processes and anthropogenic interventions. Geo-Marine Letters 33, 143-157.

Flor, G., Llera González, E.M., Martínerz Arpírez, J., Ortea Rato, J.A. 1981. Contribución al estudio de la playa de San Lorenzo (Gijón). Cuadernos del CRINAS, 1, Consejo Regional de Asturias. Consejería de Comercio, Turismo y Pesca. 47 pp. http://tematico.asturias.es/dgpesca/fich/Cuadernos%20del%20CRINAS%201.pdf

Flor-Blanco, G., Flor, G., Morales, J.A., Pando, A., 2015. Hydrodynamic controls of morpho-sedimentary evolution in a rockbounded mesotidal estuary. Tina Menor (N Spain). Journal of Iberian Geology 41, 315-332.

Flor-Blanco, G., Flor, G., Gallego, J.L., Sierra, C., Rey Díaz de Rada, J., Barranco Ojeda, A., 2016. Aspectos morfo-sedimentarios y ambientales de la playa arenosa de San Lorenzo (Gijón, Asturias, NO de España). Geo-Temas 16, 271-274.

Folk, R.L., Ward, W.C., 1957. Brazos River Bar: a study on the significance of grain size parameters. Journal of Sedimentary Petrology 27, 3-26.

Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97, 611-631.

Friedman, G.M., 1967. Dynamic processes and statistical parameters compared for size frequency distribution of beach and river sands. Journal of Sedimentary Petrology 37, 327-354.

Gaffney, S.J., 2004. Probabilistic curve-aligned clustering and prediction with regression mixture models. PhD Dissertation. University of California, Irvine. Available at: http://www.ics.uci.edu/~sgaffney/papers/sgaffney_thesis.pdf

Gaffney, S.J., Robertson, A.V., Smyth, P., Camargo, S.J., Ghil, M., 2007. Probabilistic clustering of extratropical cyclones using regression mixture models. Climate Dynamics 29, 423-440.

García, A., Sámano, M.L., Juanes, J.A., Medina, R., Revilla, J.A., Álvarez, C., 2010. Assessment of the effects of a port expansion on algae appearance in a costal bay through mathematical modelling. Application to San Lorenzo Bay (North Spain). Ecological Modelling 221, 1413-1426.

Gee, G.W., Bauder, J.W., 1996. Particle size analysis. In: Klute, A. (Ed.), Methods of soil analysis. American Society of Agronomy, Madison, WI, pp. 383-411.

Gómez, A.G., Bárcena, J.F., Juanes, J.A., Ondiviela, B., Sámano, M.L., 2014. Transport time scales as physical descriptors to characterize heavily modified water bodies near ports in coastal zones. Environmental Management 136, 76-84.

Horowitz, A.J., Elrick, K.A., 1987. The relation of stream sediment surface area, grain size and composition to trace element chemistry. Applied Geochemistry 2, 437-451.

Jackson, D., Richardson, M., 2007. High-frequency seafloor acoustics. Springer Science & Business Media, Berlin, Germany.

Knight, J., Orford, J.D., Wilson, P., Braley, S.M., 2002. Assessment of temporal changes in coastal sand dune environments using the log□hyperbolic grain□size method. Sedimentology, 49, 1229-1252.

Krumbein, W.C., 1934. Size frequency distributions of sediments. Journal of Sedimentary Petrology 4, 65–77.

McLachlan G.J., Krishnan T., 2008. The EM algorithm and extensions. Wiley, New York.

McLaren, P., Bowles, D., 1985. The effects of sediment transport on grain size distributions. Journal of Sedimentary Petrology 55, 457-470.

Ministerio de Fomento, 2010. Annual report. Port of Gijón. Gijón Port Authority. Gijón. Spain.https://www.puertogijon.es/recursos/descargas/memorias_e_informes/40657_54542011115858.pdf.

McManus, J., 1988. Grain size determination and interpretation. In: Tucker, M. (Ed.), Techniques in Sedimentology. Blackwell, UK, pp. 63-85.

Merkus, H.G., 2009. Particle size measurements: fundamentals, practice, quality. Springer, Berlin, Germany.

Passega, R., 1957. Texture as characteristic of elastic deposition. American Association of Petroleum Geologists 41, 1952-1984.

Poppe, L.J., Eliason, A.H., Fredericks, J.J., Rendigs, R.R., Blackwood, D., Polloni, C.F. 2000. Grain size analysis of marine sediments: methodology and data processing. US Geological Survey East Coast sediment analysis: procedures, database, and georeferenced displays. US Geological Survey Open File Report 00-358. http://pubs. usgs. gov/of/2000/of00-358.

Revil, A., Jardani, A., 2013. The self-potential method: theory and applications in environmental geosciences. Cambridge University Press, Cambridge, United Kingdom.

Schwarz, G.E., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461-464.

ROM 0.3-91, 1992. Oleaje. Anejo I. Clima Marítimo en el Litoral Español. MOPT. Centro de Publicaciones. Madrid.

Sahu, B.K., 1964. Depositional mechanisms from the size analysis of clastic sediments.

Journal of Sedimentary Petrology 34, 73–84.

Solohub, J.T., Klovan, J.E., 1970. Evaluation of grain-size parameters in lacustrine environments. Journal of Sedimentary Petrology 40, 81–101.

Syvitski, J.P., 2007. Principles, methods and application of particle size analysis. Cambridge University Press, Cambridge, United Kingdom.

Tibshirani, R., Walther, G., Hastie. T., 2001. Estimating the number of data clusters via the gap statistic. Journal of the Royal Statistical Society, Series B, Statistical Methodology 63, 411-423.

Vandenberghe, J., 2013. Grain size of fine-grained windblown sediment: A powerful proxy for process identification. Earth-Science Reviews 121, 18-30.

Vega, L., Medina, R., 2010. Informe sobre el estado actual y evolución de la playa de San Lorenzo tras las obras de ampliación del puerto de Gijón y propuestas de regeneración. Informe Final. Instituto de Hidráulica Ambiental (IH). Universidad de Cantabria. Autoridad Portuaria de Gijón. Gijón, Spain. https://www.puertogijon.es/recursos/descargas/medio__ambiente/playa_de_san_lorenzo/46944_271027102010181949.pdf.

Weltje, G.J., 1997. End-member modeling of compositional data: Numerical-statistical algorithms for solving the explicit mixing problem. Mathematical Geology 29, 503-549.

Weltje, G.J., Prins, M.A., 2007. Genetically meaningful decomposition of grain-size distributions. Sedimentary Geology 202, 409-424.

Weltje, G.J., von Eynatten, H., 2004. Quantitative provenance analysis of sediments: review and outlook. Sedimentary Geology 171, 1-11.

Figure legends

Figure 1. Location of Gijon in Asturias (north of Spain) and spatial distribution of the sampling dots on an orthophoto map.

Figure 2. Initial grain-size probability density functions obtained after laser diffraction spectroscopy.

Figure 3. First four smoothed functional principal components (FPC1-4) for the grain-size data. The first two functional components explain 94.5% of the total variation.

Figure 4. Probability density curves with significantly positive (continuous line) and negative (discontinuous line) scores of FPC1 (a) and FPC2 (b).

Figure 5. Determination of the knee point of Bayesian information criterion (BIC).

Figure 6. Grain-size probability density functions grouped on the basis of three clusters (cluster 1: red, cluster 2: deep blue, cluster 3: green).

Figure 7. Spatial distribution of the three clusters obtained by FCA (cluster 1: red, cluster 2: deep blue, cluster 3: green).

Figure 8. Grain-size distributions assigned to each of the 3 clusters obtained using PDF approach (a) and PCA-CA approach on clr-transformed data (b).

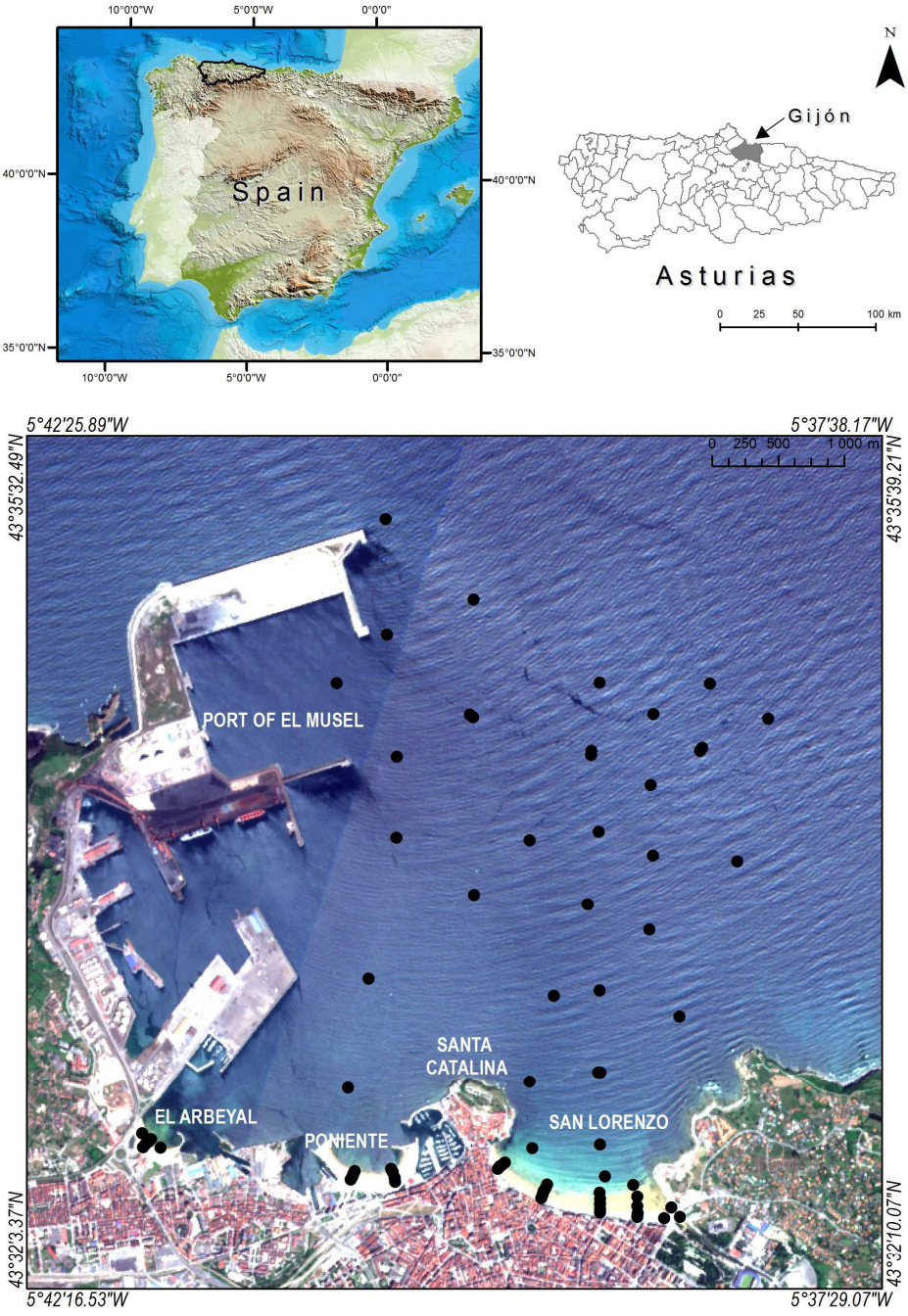

Figure 1

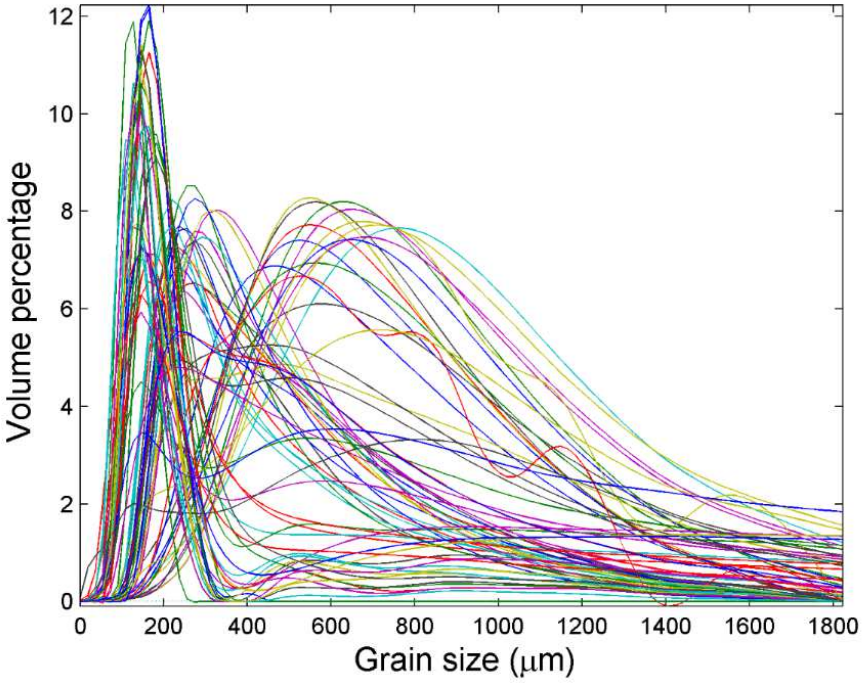

Figure 2

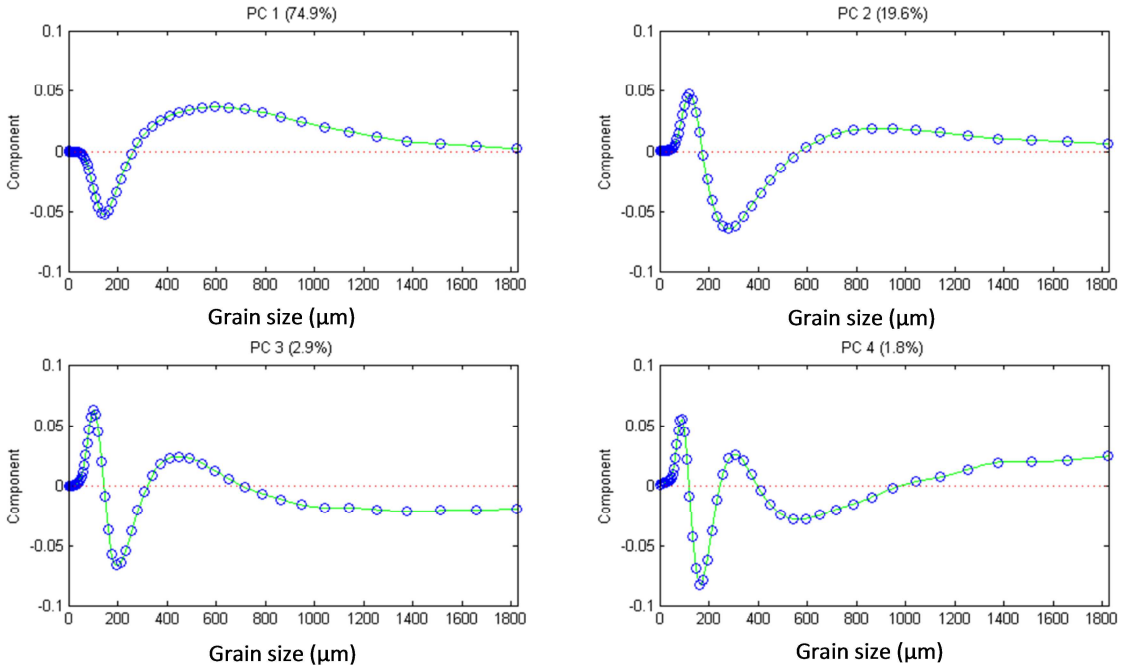

Figure 3

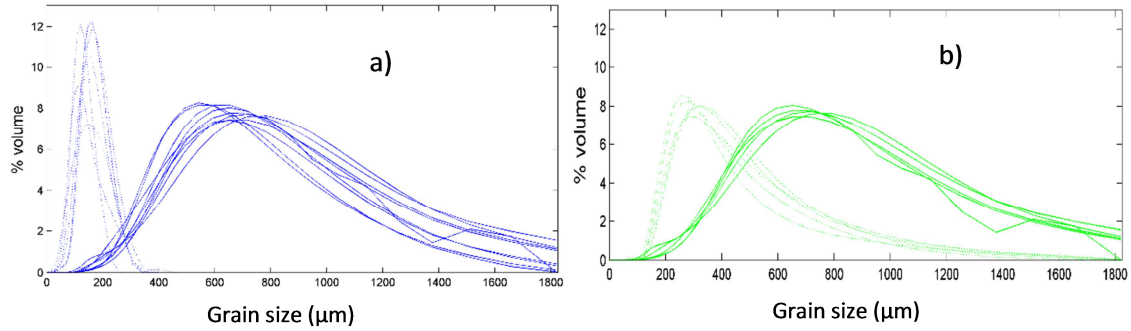

Figure 4

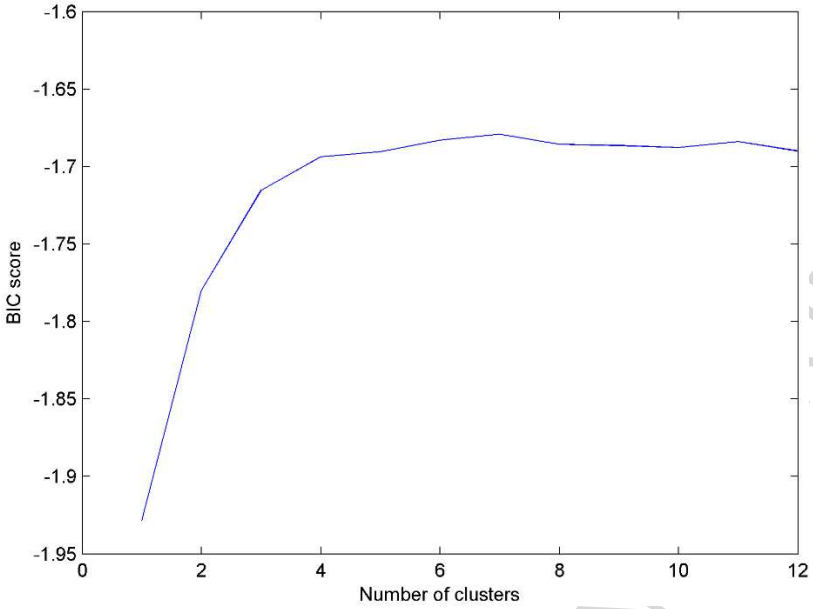

Figure 5

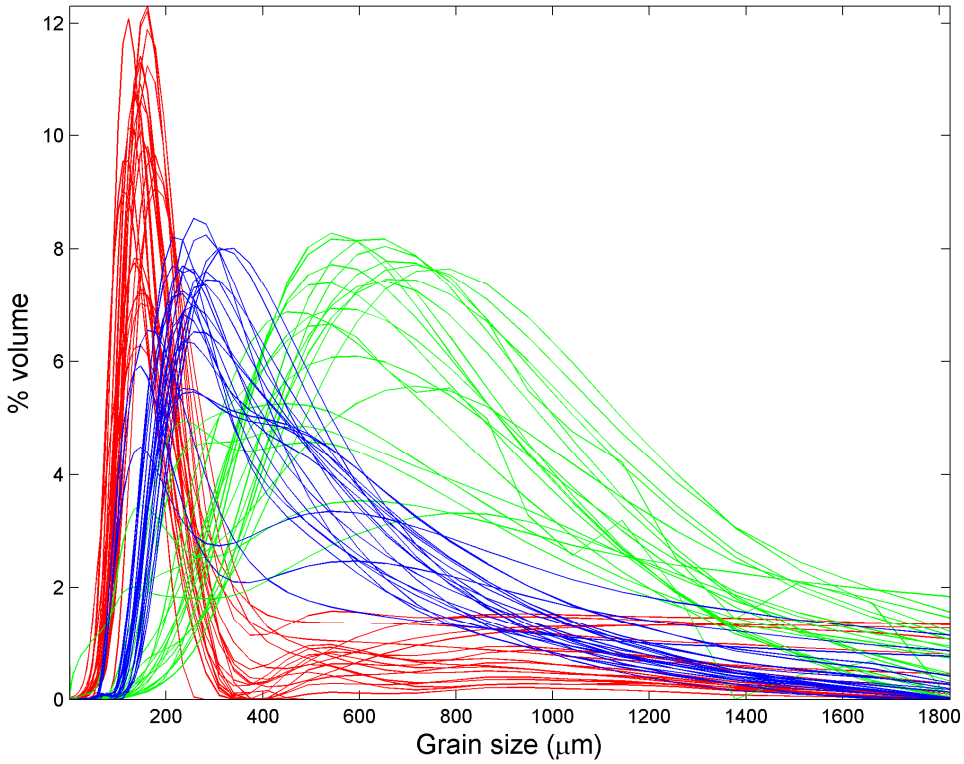

Figure 6

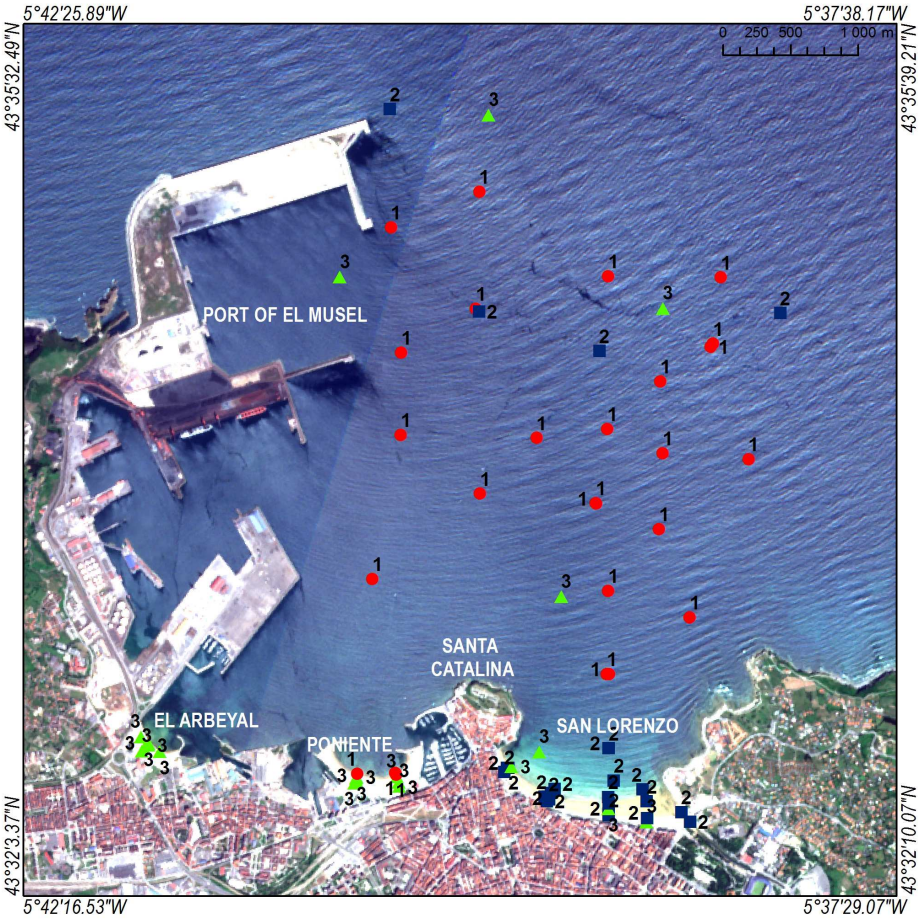

Figure 7

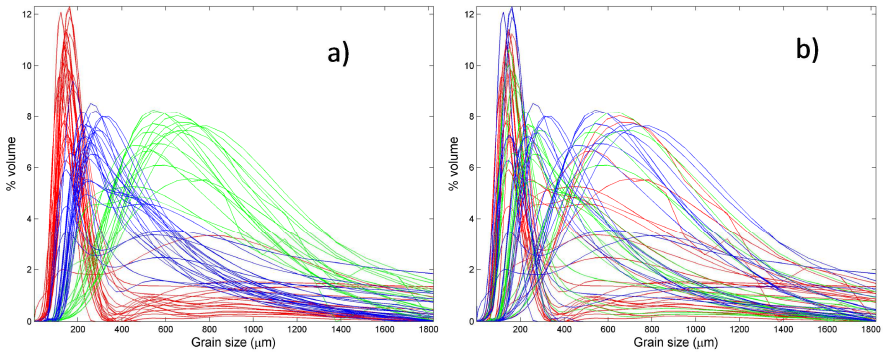

Figure 8

Table 1. Average parameters obtained by the arithmetic method of moments of the curves with positive and negative values for the first two functional principal components (FPCs).

Parameter	Unit	FPC 1		FPC 2	
		positive	negative	positive	negative
Mean	µm	716.1	148.1	747.8	382.9
D_{10}		341.4	75.9	335.1	192.2
D_{50}		663.3	142.1	704.9	318.4
D_{90}		1.175.7	230.9	1.259.2	653.8
Sorting		330.6	60.2	356.3	219.7
Skewness	-	0.8	0.3	0.6	2.1
Kurtosis	-	3.4	2.9	3.1	8.9

Table 2. Statistical descriptors (arithmetic method) and carbonate concentrations of each group for a three-cluster apportionment.

Parameter	Unit	Cluster 1	Cluster 2	Cluster 3
Mean	µm	254.9	425.3	665.6
D_{10}		106.5	187.1	248.6
D_{50}		169.9	346.8	616.2
D_{90}		597.7	783.1	1219.5
Sorting		244.4	261.5	367.1
Skewness	-	4.4	1.7	0.7
Kurtosis	-	32.2	6.5	3.1
Carbonates	%	34.3	59.2	76.4

Table 3. Loadings of each log-normal parameters

	PC1	PC2	PC3	PC4
Mean	0.9992	-0.039	0.0006	0.0008
Sorting	0.0000	0.006	-0.5869	0.8096
Skewness	-0.0062	-0.1315	0.8021	0.5825
Kurtosis	0.0390	0.9905	0.1102	0.0722